Building CoronaWhy Knowledge Graph

FREYA Guest webinar

Slava Tykhonov
Senior Information Scientist
(DANS-KNAW, the Netherlands)

01.09.2020

About me: DANS-KNAW projects (2016-2020)

- CLARIAH+ (ongoing)
- EOSC Synergy (ongoing)
- SSHOC Dataverse (ongoing)
- CESSDA DataverseEU 2018
- Time Machine Europe Supervisor at DANS-KNAW
- PARTHENOS Horizon 2020
- CESSDA PID (Personal Identifiers) Horizon 2020
- CLARIAH
- RDA (Research Data Alliance) PITTS Horizon 2020
- CESSDA SaW H2020-EU.1.4.1.1 Horizon 2020

Source: LinkedIn

Motivation

Slava Tykhonov 4:05 PM

Born in Kyiv, Ukraine but raised as a scientist in the Netherlands, during my career I saw a lot of cases where people refused to collaborate and work together due to own ambitions and wrong vision. I truly believe that dangerous things like coronavirus are possible in the modern society only because people are competing against each other and don't want to share their knowledge and competence, and work together in order to find a solution for a problem.

Stupidity, ignorance and limitness has no nationality, it's just a common thing that killing this world and COVID-19 is just one of the challenges. I'm here to open everything that should be open for the humanity, build the collaboration between people, speed up the innovation and start the development of the research infrastructure that will allow to bring Science back to the policy table, and quickly respond to the current and future challenges. I would say, we should be prepared for the technological future and don't afraid to disrupt the world. (edited)

👍 6 💪 1 ❤️ 3

7 weeks in lockdown in Spain

[Resistere (I will resist)]

About CoronaWhy

www.coronawhy.org

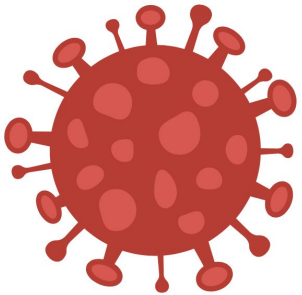
CORONAWHY

Home Daily Progress Calendar JOIN THE FIGHT!

FIGHTING CORONAVIRUS
WITH ARTIFICIAL INTELLIGENCE

We are a globally distributed, volunteer-powered research organisation, trying to assist the medical community's ability to answer key questions related to COVID-19

JOIN THE FIGHT LEARN MORE



Who we are?

Artur Kiulian started CoronaWhy because he realized that we are all in this together. Now CoronaWhy is an international group of **900+ volunteers** whose mission is to improve global coordination and analysis of all available data pertinent to the COVID-19 outbreak and ensure all findings reach those who need them.

It's impossible to list everyone out but we will eventually.

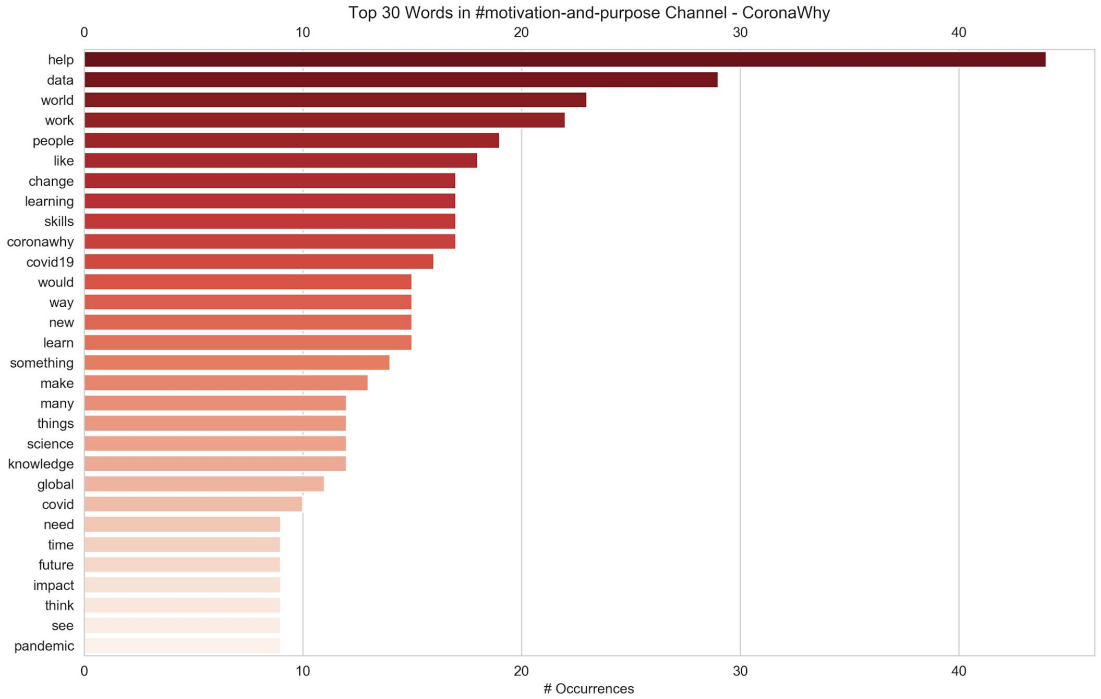
1300+ people registered in the organization, more than 300 actively contributing!

COVID-19 Open Research Dataset Challenge (CORD-19)

It's all started from this (March, 2020):

"In response to the COVID-19 pandemic and with the view to boost research, the Allen Institute for AI together with CZI, MSR, Georgetown, NIH & The White House is collecting and making available for free the COVID-19 Open Research Dataset (CORD-19). This resource is updated weekly and contains over 52,000 scholarly articles, including 41,000 with full text, about COVID-19 and other viruses of the coronavirus family." (Kaggle)

Motivation of CoronaWhy community members

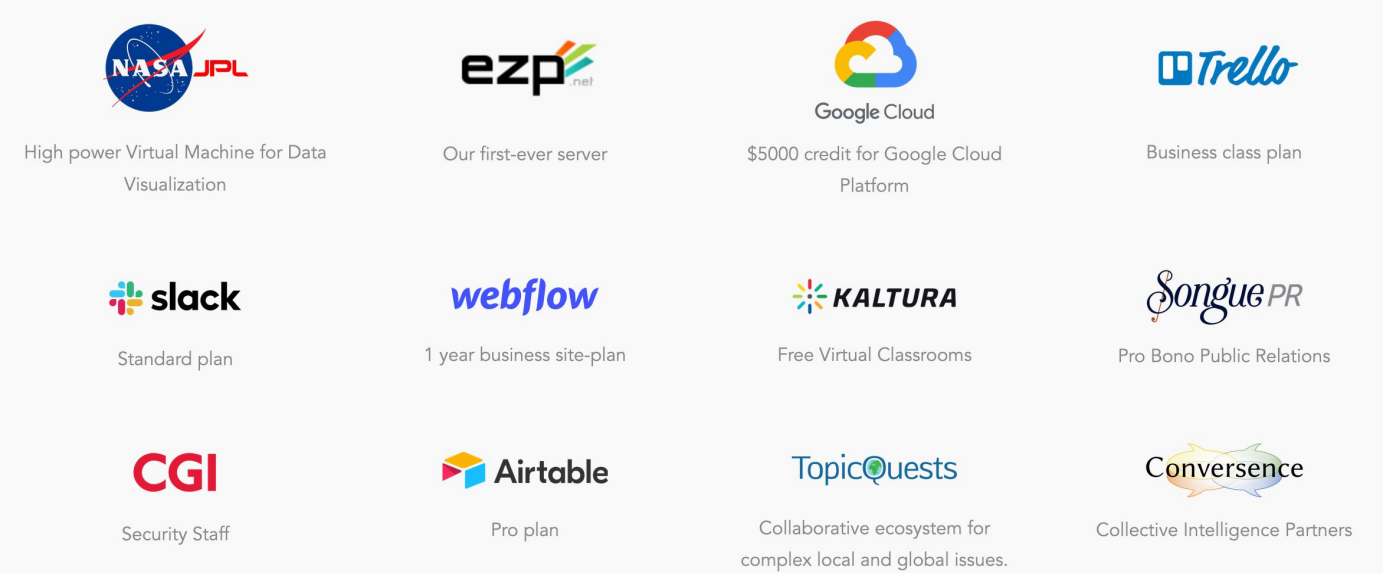


Credits: Andre Ye

CoronaWhy Funding

Initial: $5k from Google on GCP and $4k from Amazon on AWS (April 2020)

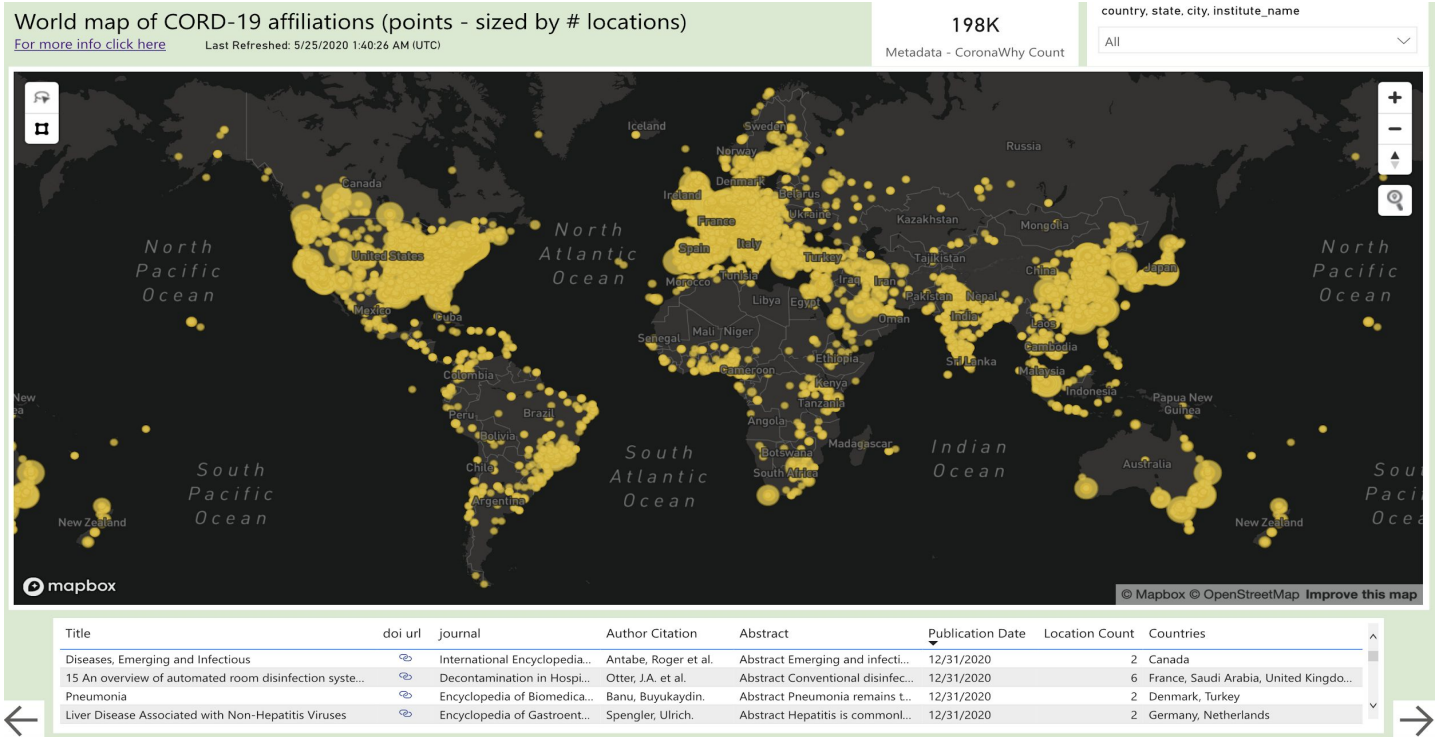
Donations: $9k and 15k british pounds to sustain CoronaWhy infrastructure



CoronaWhy Community Tasks (March-April)

1. Task-Risk helps to identify risk factors that can increase the chance of being infected, or affects the severity or the survival outcome of the infection
2. Task-Ties to explore transmission, incubation and environment stability
3. Match Clinical Trials allows exploration of the results from the COVID-19 International Clinical Trials dataset
4. COVID-19 Literature Visualization helps to explore the data behind the AI-powered literature review
5. Named Entity Recognition across the entire corpus of CORD-19 papers with full text

CORD-19 affiliations recognized with Deep Learning



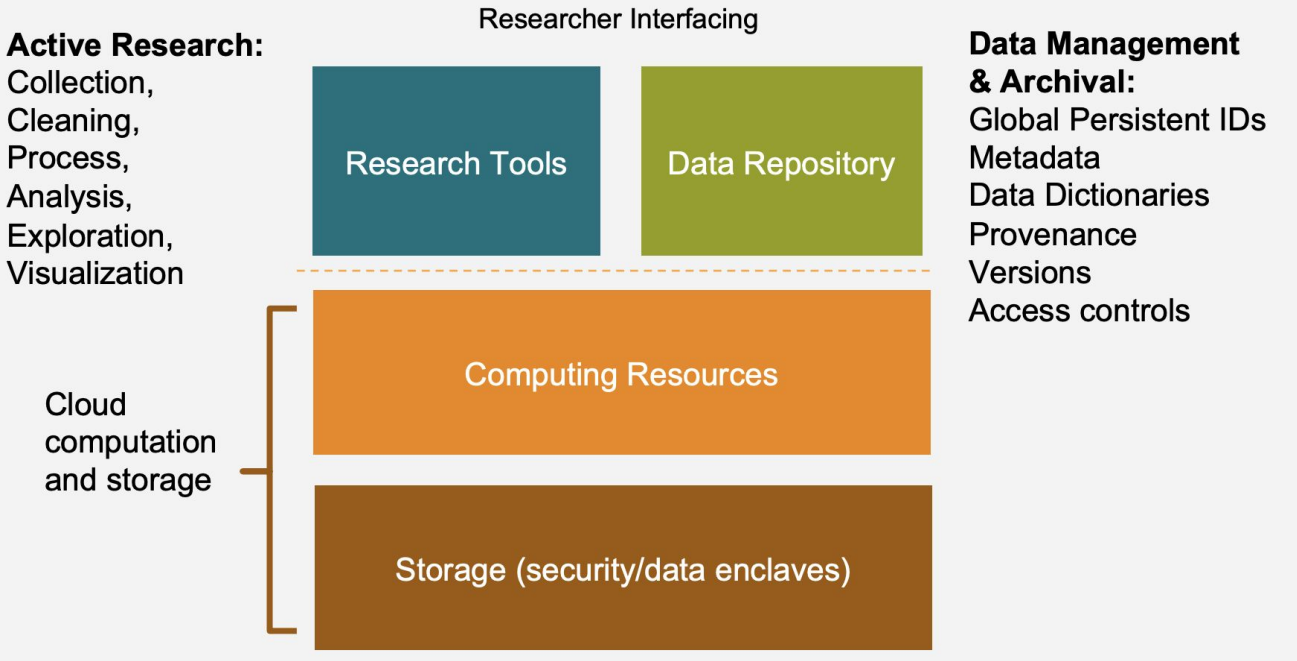
Source: [CORD-19 map visualization] and [institution affiliation data]

Collaboration with other organizations

- Harvard Medical School, INDRA integration
- Helix Group, Stanford University
- NASA JPL, COVID-19 knowledge graph and GeoParser
- Kaggle, coronamed application
- Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS, knowledge graph
- dcyphr, a platform for creating and engaging with distillations of academic articles
- CAMARADES (Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies)

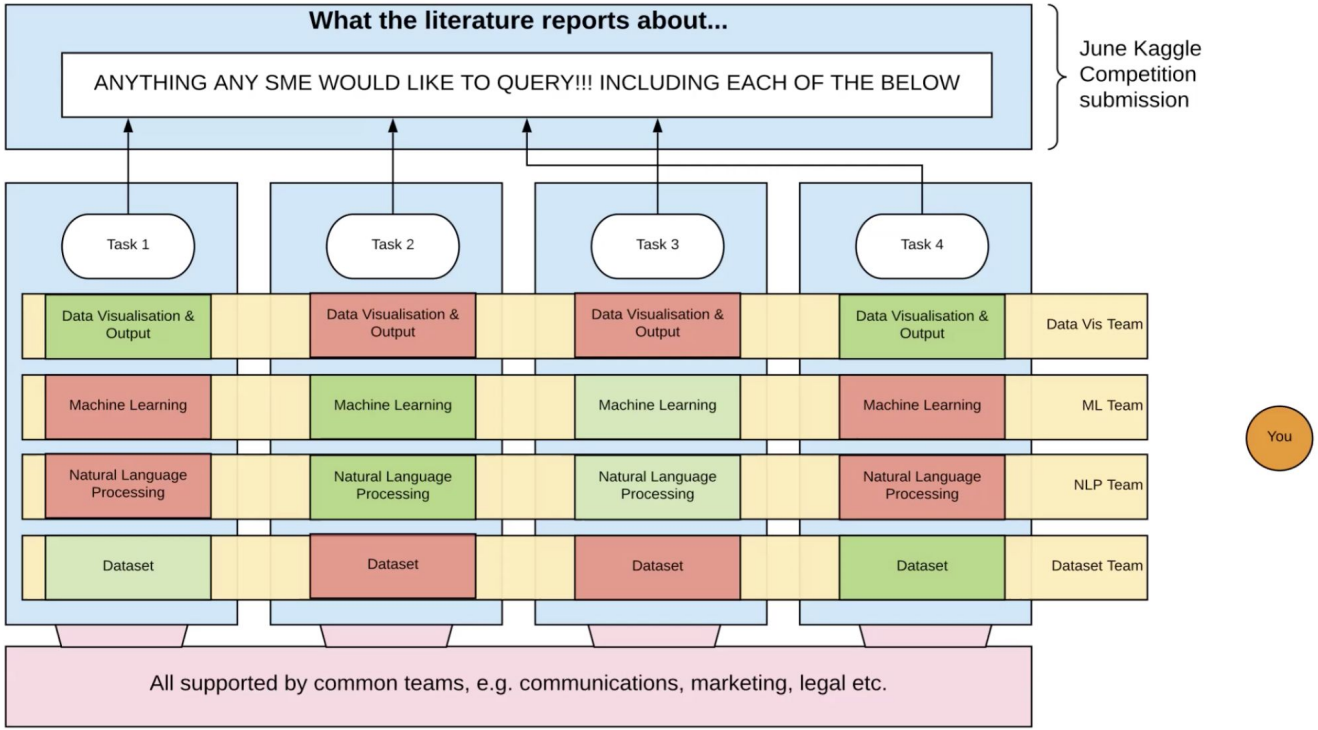
We’ve got almost endless data streams...

Looking for Commons



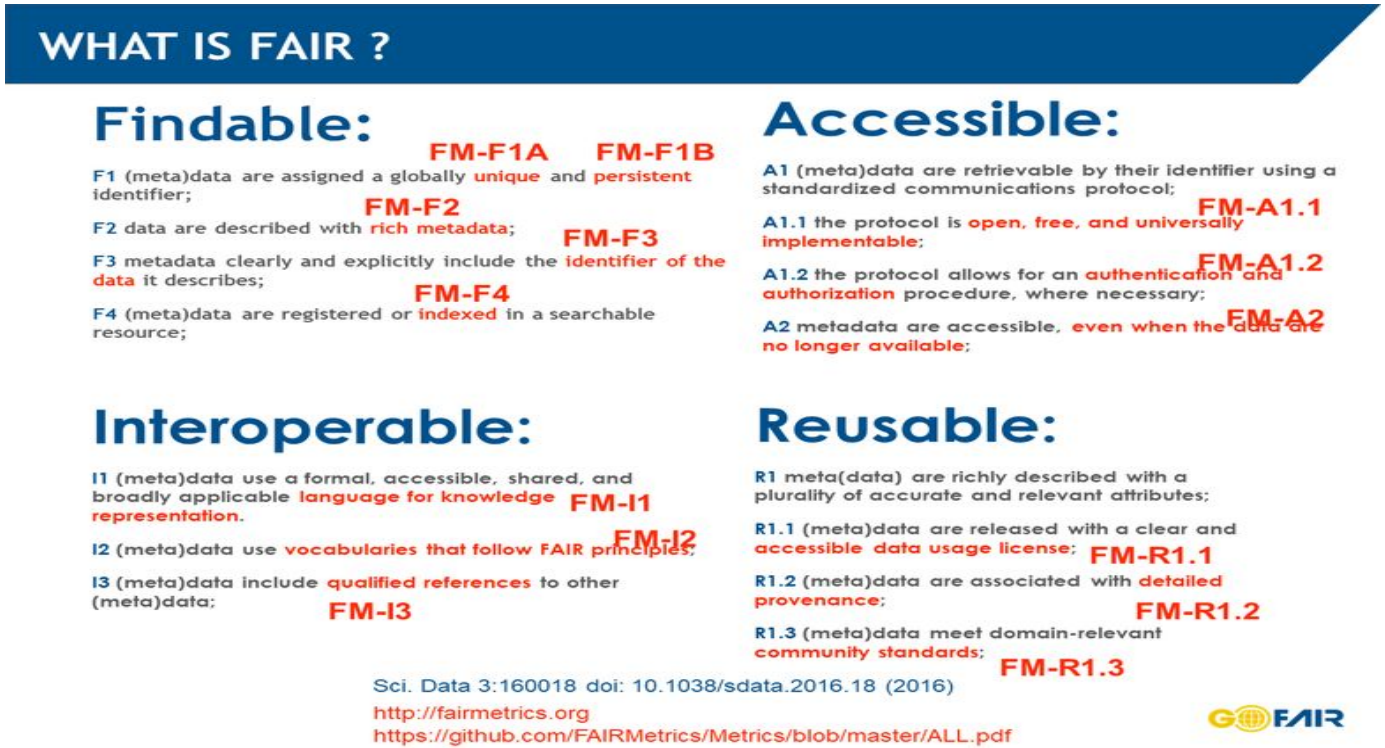
Merce Crosas, "Harvard Data Commons"

Building a horizontal platform to serve vertical teams



Source: [CoronaWhy infrastructure](#) introduction

Turning FAIR into reality!



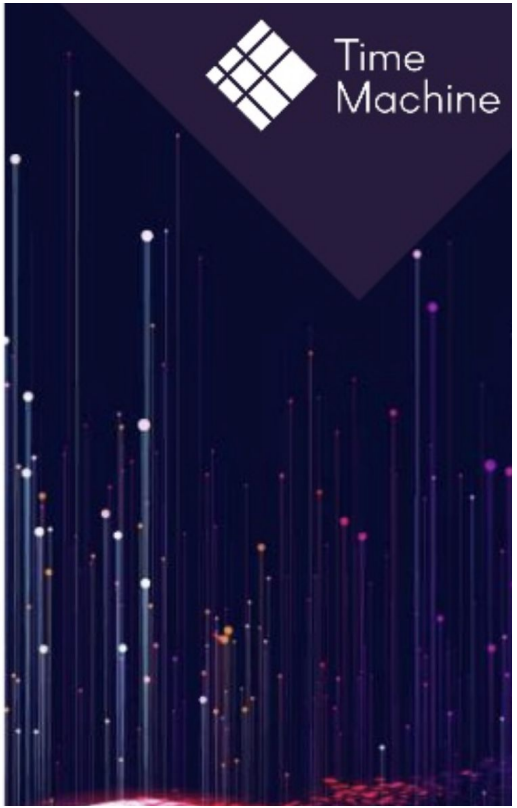
DANS-KNAW is one of worldwide leaders in FAIR Data (FAIRsFAIR)

Standing on the Shoulders of Giants: infrastructure

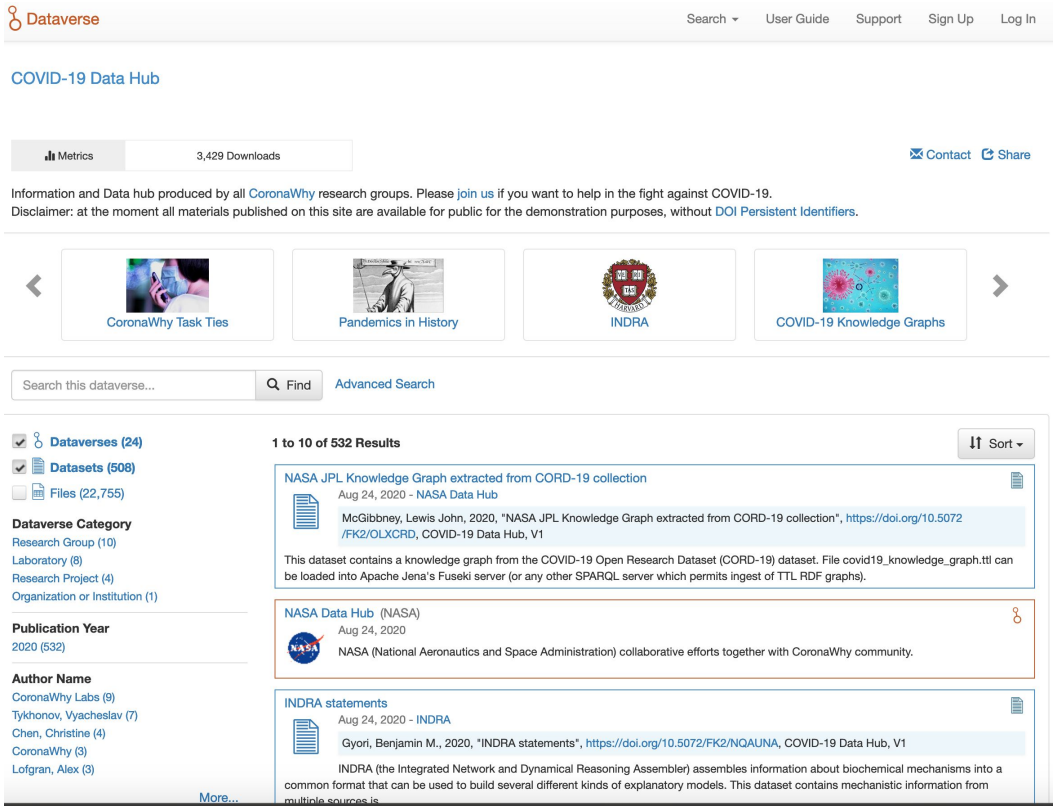
Standing on the Shoulders of Giants: Big Data of the Past

Time Machine is …

- An international **collaboration** to bring 5000 years of European history to life
- Digitising millions of **historical documents**, painting and monuments
- The **largest** computer simulation ever developed
- An **open access**, interactive resource



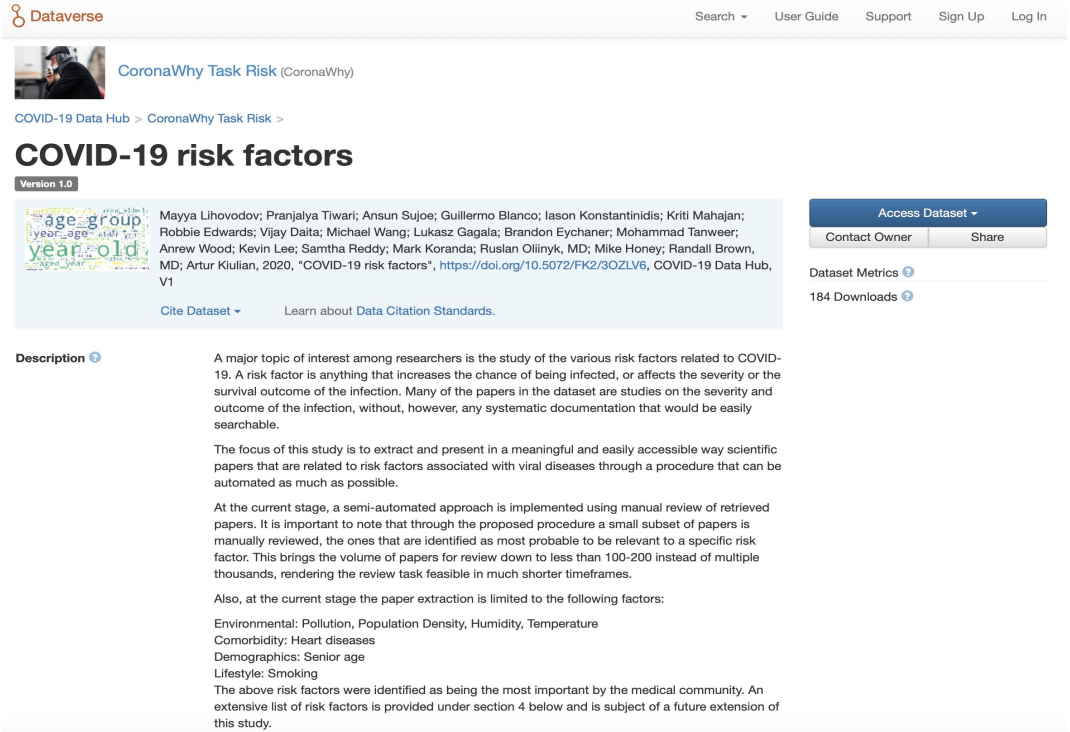
Dataverse as data integration point



- Available as a service for the community from April, 2020
- Used by CoronaWhy vertical teams for the data exchange and share
- Intended to help researchers to make their data FAIR
- One of the biggest COVID-19 data archives in the world with 700k files
- New teams are getting own data containers and can reuse data collected and produced by others

http://datasets.coronawhy.org

Dataset from CoronaWhy vertical teams



Dataverse

Search ▾ User Guide Support Sign Up Log In

CoronaWhy Task Risk (CoronaWhy)

COVID-19 Data Hub > CoronaWhy Task Risk >

COVID-19 risk factors

Version 1.0

Mayya Lihovodov; Pranjalya Tiwari; Ansun Sujoe; Guillermo Blanco; Iason Konstantinidis; Kriti Mahajan; Robbie Edwards; Vijay Daita; Michael Wang; Lukasz Gagala; Brandon Eychaner; Mohammad Tanweer; Anrew Wood; Kevin Lee; Samtha Reddy; Mark Koranda; Ruslan Oliinyk, MD; Mike Honey; Randall Brown, MD; Artur Kiulian, 2020, "COVID-19 risk factors", https://doi.org/10.5072/FK2/3OZLV6, COVID-19 Data Hub, V1

Cite Dataset ▾ Learn about Data Citation Standards.

Access Dataset ▾

Contact Owner Share

Dataset Metrics

184 Downloads

Description

A major topic of interest among researchers is the study of the various risk factors related to COVID-19. A risk factor is anything that increases the chance of being infected, or affects the severity or the survival outcome of the infection. Many of the papers in the dataset are studies on the severity and outcome of the infection, without, however, any systematic documentation that would be easily searchable.

The focus of this study is to extract and present in a meaningful and easily accessible way scientific papers that are related to risk factors associated with viral diseases through a procedure that can be automated as much as possible.

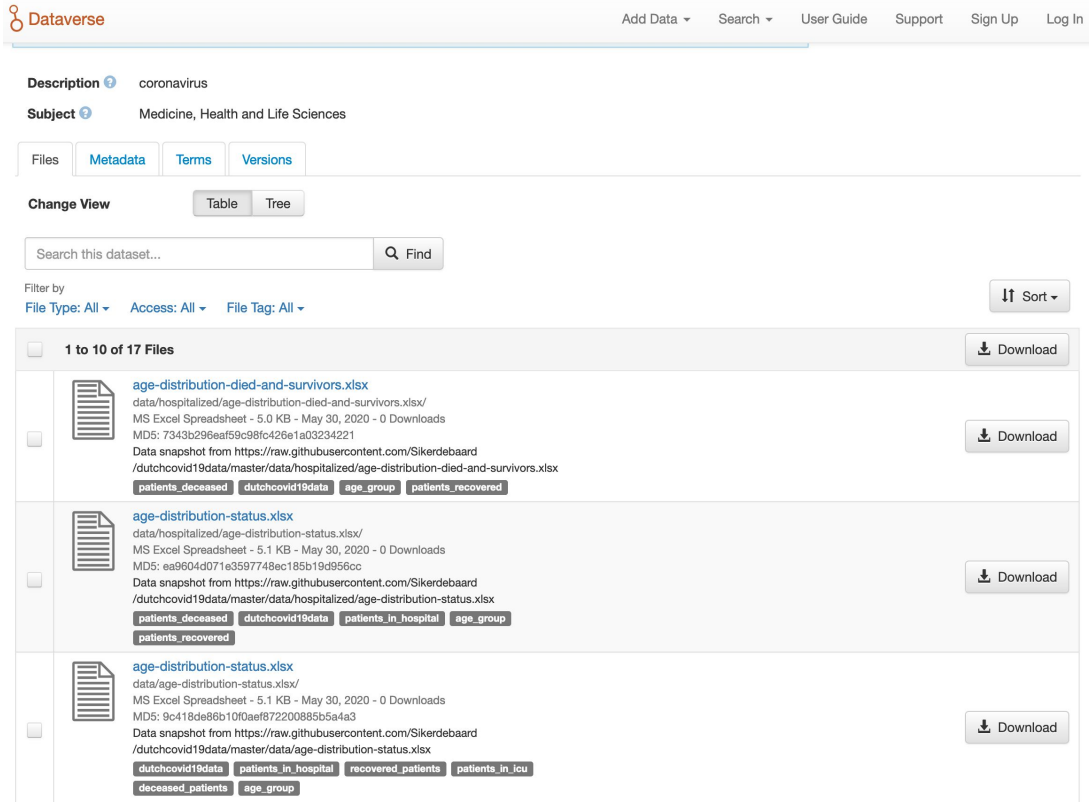
At the current stage, a semi-automated approach is implemented using manual review of retrieved papers. It is important to note that through the proposed procedure a small subset of papers is manually reviewed, the ones that are identified as most probable to be relevant to a specific risk factor. This brings the volume of papers for review down to less than 100-200 instead of multiple thousands, rendering the review task feasible in much shorter timeframes.

Also, at the current stage the paper extraction is limited to the following factors:

Environmental: Pollution, Population Density, Humidity, Temperature
Comorbidity: Heart diseases
Demographics: Senior age
Lifestyle: Smoking
The above risk factors were identified as being the most important by the medical community. An extensive list of risk factors is provided under section 4 below and is subject of a future extension of this study.

Source: [CoronaWhy Dataverse]

COVID-19 data files verification



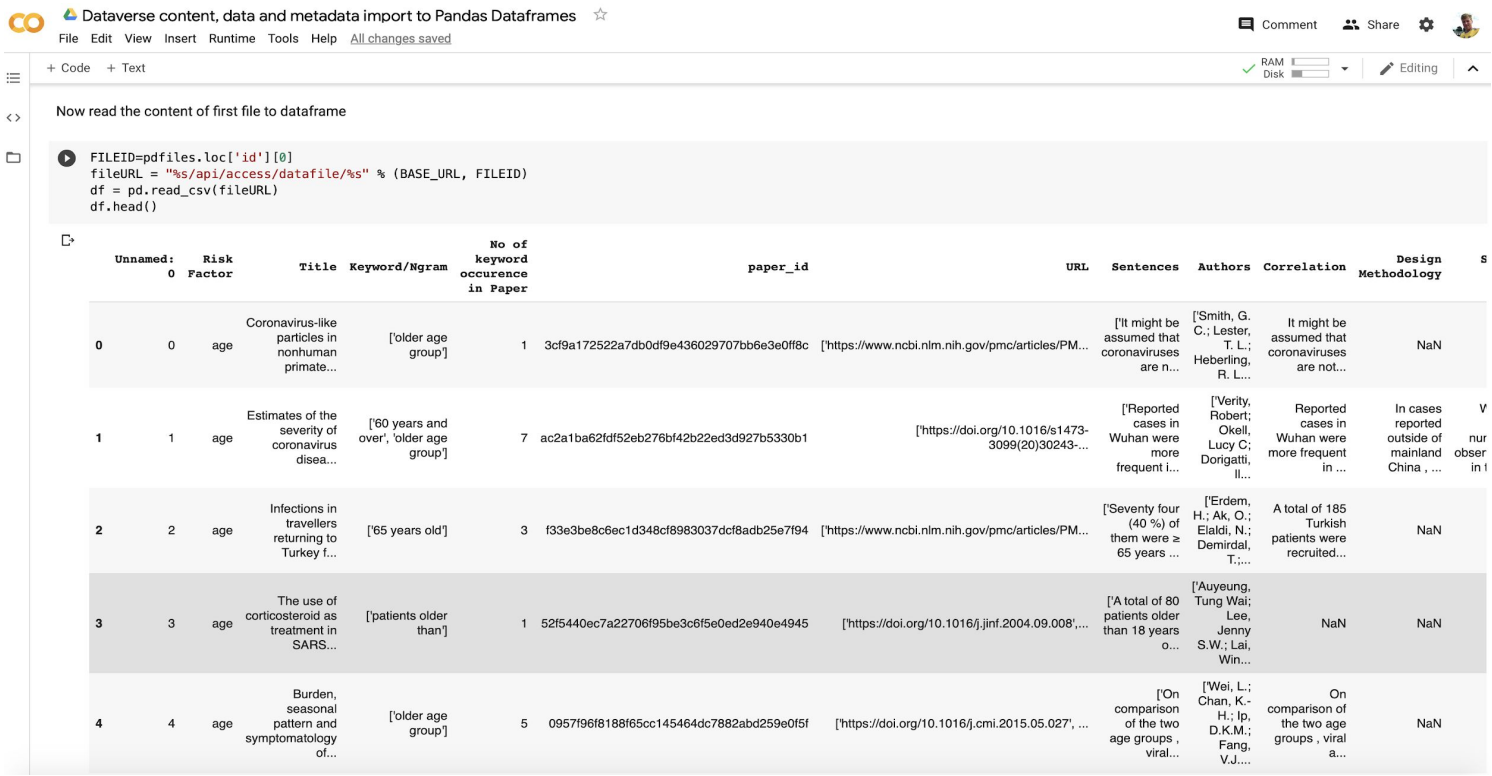
We do a verification of every file by importing its contents to dataframe.

All column names (variables) extracted from tabular data available as labels in files metadata

We've enabled Dataverse data previewers to browse through the content of files without download!

We're starting internal challenges to build ML models for the metadata classification

Dataverse content in Jupyter notebooks



Dataverse content, data and metadata import to Pandas Dataframes

File Edit View Insert Runtime Tools Help All changes saved

Comment Share Editing

+ Code + Text

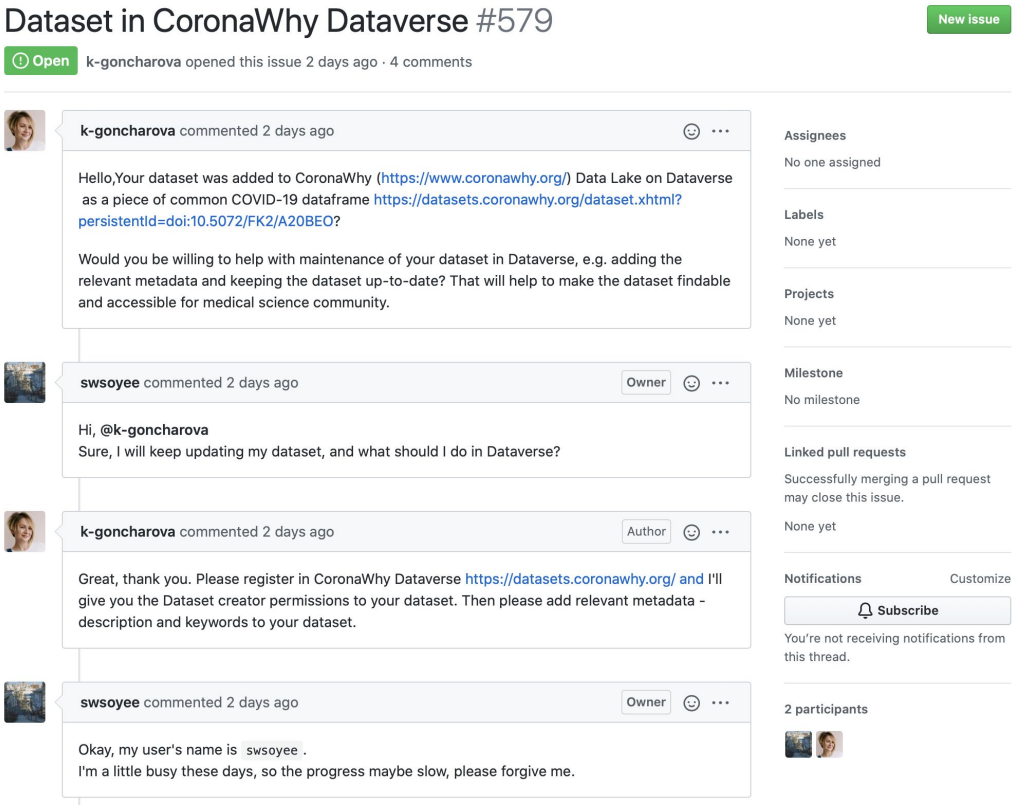
Now read the content of first file to dataframe

```
FILEID=pdfiles.loc['id'][0]
fileURL = "%s/api/access/datafile/%s" % (BASE_URL, FILEID)
df = pd.read_csv(fileURL)
df.head()
```

	Unnamed: 0	Risk Factor	Title	Keyword/Ngram	No of keyword occurence in Paper	paper_id	URL	Sentences	Authors	Correlation	Design Methodology	S
0	0	age	Coronavirus-like particles in nonhuman primate...	['older age group']	1	3cf9a172522a7db0df9e436029707bb6e3e0ff8c	['https://www.ncbi.nlm.nih.gov/pmc/articles/PM...	['It might be assumed that coronaviruses are n...	['Smith, G. C.; Lester, T. L.; Heberling, R. L....	It might be assumed that coronaviruses are not...	NaN	
1	1	age	Estimates of the severity of coronavirus disea...	['60 years and over', 'older age group']	7	ac2a1ba62fdf52eb276bf42b22ed3d927b5330b1	['https://doi.org/10.1016/s1473-3099(20)30243-...	['Reported cases in Wuhan were more frequent i...	['Verity, Robert; Okell, Lucy C; Dorigatti, Il...	Reported cases in Wuhan were more frequent in ...	In cases reported outside of mainland China , ...	V nur obser in t
2	2	age	Infections in travellers returning to Turkey f...	['65 years old']	3	f33e3be8c6ec1d348cf8983037dcf8adb25e7f94	['https://www.ncbi.nlm.nih.gov/pmc/articles/PM...	['Seventy four (40 %) of them were ≥ 65 years ...	['Erdem, H.; Ak, O.; Elaldi, N.; Demirdal, T.;...	A total of 185 Turkish patients were recruited...	NaN	
3	3	age	The use of corticosteroid as treatment in SARS...	['patients older than']	1	52f5440ec7a22706f95be3c6f5e0ed2e940e4945	['https://doi.org/10.1016/j.jinf.2004.09.008',...	['A total of 80 patients older than 18 years o...	['Auyeung, Tung Wai; Lee, Jenny S.W.; Lai, Win...	NaN	NaN	
4	4	age	Burden, seasonal pattern and symptomatology of...	['older age group']	5	0957f96f8188f65cc145464dc7882abd259e0f5f	['https://doi.org/10.1016/j.cmi.2015.05.027', ...	['On comparison of the two age groups , viral...	['Wei, L.; Chan, K.-H.; Ip, D.K.M.; Fang, V.J....	On comparison of the two age groups , viral a...	NaN	

Source: [Dataverse examples on Google Colabs]

COVID-19 Data Crowdsourcing



CoronaWhy data management team does does the review of all harvested datasets and try to identify the important data.

We're approaching github owners by creating issues in their repos and inviting them to help us.

More than 20% of data owners joining CoronaWhy community or interested to curate their datasets.

Bottom-up data collection works!

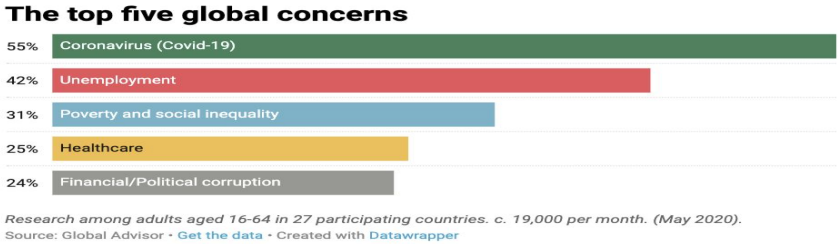
Challenge of data integration and various ontologies

CORD-19 collection processing with NLP pipeline:

- manual **annotation** and **labelling** of COVID-19 related papers
- automatic **entity extraction** and **classification** of text fragments
- **statements** extraction and curation
- linking papers to specific research questions with **relationships** extraction

Dataverse Data Lake streaming COVID-19 datasets from various sources:

- medical data
- socio-economic Data
- political data and statistics



The importance of standards and ontologies

Generic controlled vocabularies to link metadata in the bibliographic collections are well known: ORCID, GRID, GeoNames, Getty

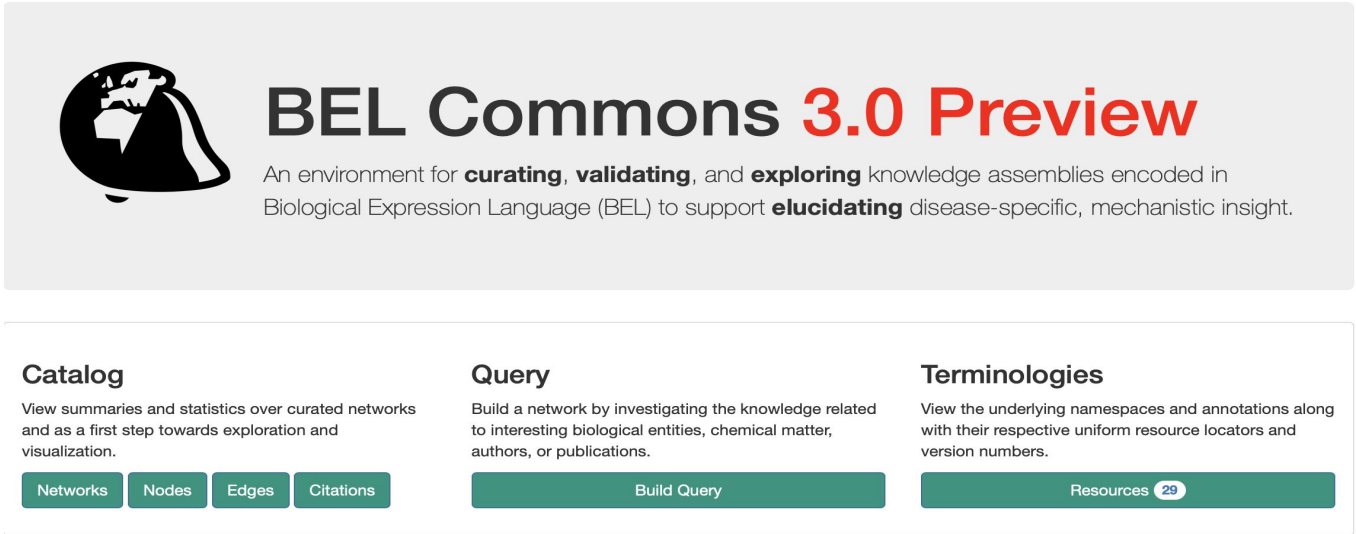
Medical knowledge graphs powered by:

- Biological Expression Language (BEL)
- Medical Subject Headings (MeSH®) by U.S. National Library of Medicine (NIH)
- Wikidata (Open ontology) - Wikipedia

Integration based on metadata standards:

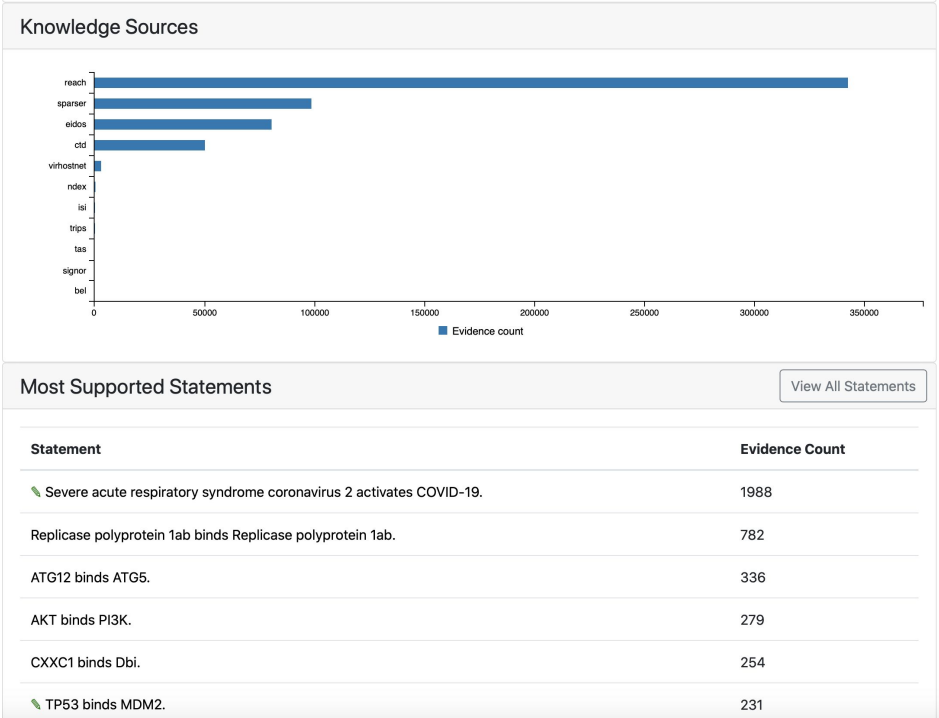
- MARC21, Dublin Core (DC), Data Documentation Initiative (DDI)

Biological Expression Language (BEL)



BEL was integrated in CoronaWhy infrastructure in April, 2020

Statements extraction with INDRA



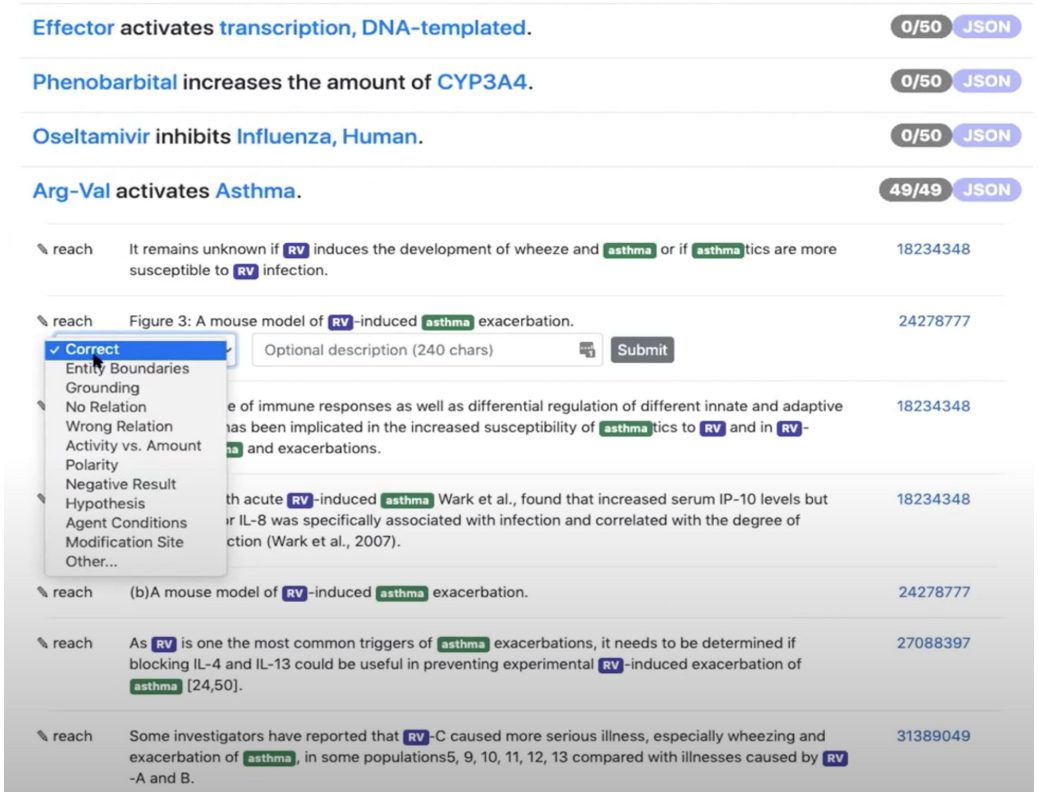
"INDRA (Integrated Network and Dynamical Reasoning Assembler) is an automated model assembly system, originally developed for molecular systems biology and currently being generalized to other domains."

Developed as a part of Harvard Program in Therapeutic Science and the Laboratory of Systems Pharmacology at Harvard Medical School.

http://indra.bio

Source: EMMAA (Ecosystem of Machine-maintained Models with Automated Assembly)

Knowledge Graph curation in INDRA

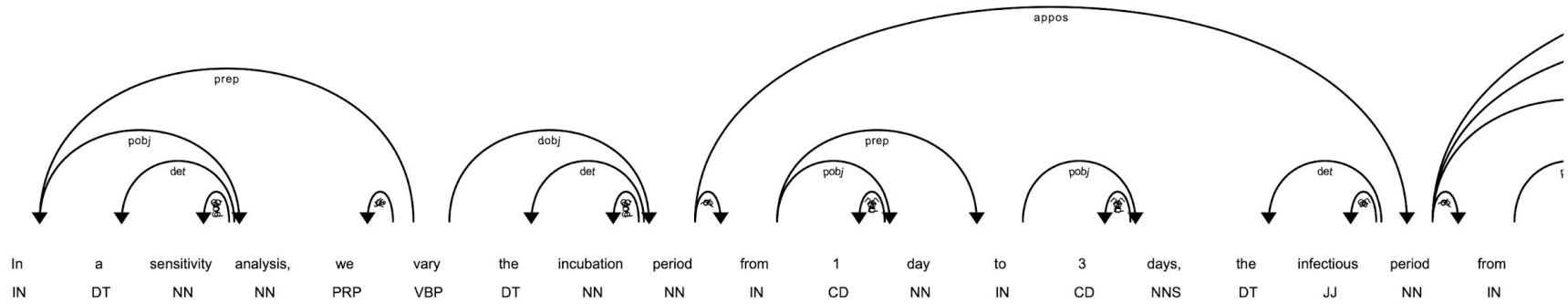


Advanced Text Analysis (NLP pipeline)

We need a good understanding of all domain specific texts to make right statements:

```
[ ]  from spacy import displacy

     # 'From our definition of incubation period, his incubation period will be 28 days, which is surprisingly high.'
     doc = nlp(t[0])
     displacy.render(doc, style='dep', jupyter=True, options={'distance': 80, "fine_grained": True})
```



Source: D.Shlepakov, [How to Build Knowledge Graph, notebook]

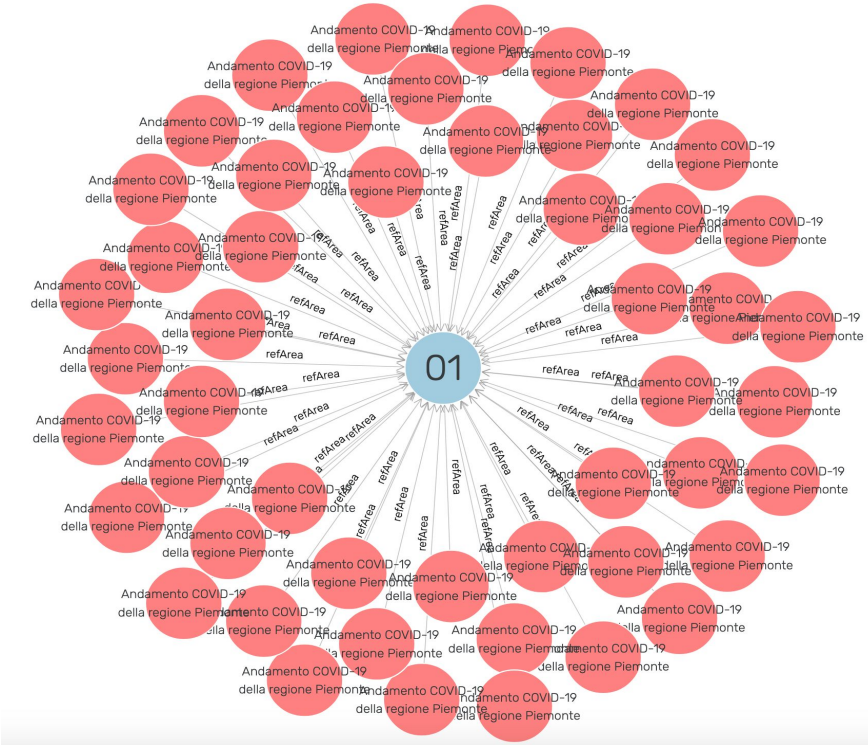
Building domain specific knowledge graphs

- We’re collecting all possible COVID-19 data and archiving in our Dataverse
- Looking for various related controlled vocabularies and ontologies
- Building and reusing conversion pipelines to get all data values linked in RDF format

```
<http://www.protezionecivile.gov.it/dataset/covid19/national-trend/observations/20200224> a qb:Observation ;
    rdfs:label "Andamento nazionale di giorno 24/02/2020"^^xsd:string ;
    qb:dataset <http://www.protezionecivile.gov.it/dataset/covid19/national-trend> ;
    sdmx-dimension:refArea <https://w3id.org/italia/controlled-vocabulary/territorial-classifications/countries/italy/ITA> ;
    sdmx-dimension:refTime "2020-02-24"^^xsd:date ;
    dpc:deads "7"^^xsd:int ;
    dpc:healed "1"^^xsd:int ;
    dpc:homeIsolation "94"^^xsd:int ;
    dpc:hospitalizedWithSymptoms "101"^^xsd:int ;
    dpc:intensiveCare "26"^^xsd:int ;
    dpc:newPositive "221"^^xsd:int ;
    dpc:swabs "4324"^^xsd:int ;
    dpc:totalCases "229"^^xsd:int ;
    dpc:totalHospitalized "127"^^xsd:int ;
    dpc:totalPositive "221"^^xsd:int ;
    dpc:totalPositiveVariation "0"^^xsd:int .
```

The ultimate goal is to automate the process of the Knowledge extraction by using the latest developments in Artificial Intelligence and Deep Learning.

Visual graph of COVID-19 dataset



Andamento COVID-19 della regione Piemonte

Andamento COVID-19 della regione Piemonte

Types:

qb:Observation

RDF rank:

0

Search instance properties

sdmx-dimension:refTime
2020-03-12

dpc:deads
26

dpc:healed
0

dpc:homeIsolation
89

dpc:hospitalizedWithSymptoms
368

dpc:intensiveCare
97

dpc:newPositive
79

dpc:swabs
2879

Source: CoronaWhy GraphDB

SPARQL endpoint for CoronaWhy KG

Query

https://sparql.labs.coronawhy.org/sparql

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX mesh: <http://id.nlm.nih.gov/mesh/>
PREFIX grid: <https://www.grid.ac/institutes/>
PREFIX bf: <http://id.loc.gov/ontologies/bibframe/>
PREFIX loc: <http://id.loc.gov/ontologies/bibframe/source>
PREFIX mads: <http://www.loc.gov/mads/rdf/v1#>

select * from <http://coronawhy.org/> where
{
?cord rdfs:label "United States" .
mesh:C0243052 ?a ?b
}
```

Table | Response | Gallery | Chart | Geo | Geo-3D | Geo events | Pivot | Timeline | 10000 results in 3.051 seconds | Filter query results | Page size: 50

	cord	a	b
1	<http://vufind.apps.coronawhy.org/vufind/Record/03bwk538#Place651-7>	rdf:type	bf:Topic
2	<http://vufind.apps.coronawhy.org/vufind/Record/094d0rn6#Place651-7>	rdf:type	bf:Topic
3	<http://vufind.apps.coronawhy.org/vufind/Record/2egeyh0j#Place651-7>	rdf:type	bf:Topic
4	<http://vufind.apps.coronawhy.org/vufind/Record/3p2rqavh#Place651-7>	rdf:type	bf:Topic
5	<http://vufind.apps.coronawhy.org/vufind/Record/5kapn32k#Place651-7>	rdf:type	bf:Topic
6	<http://vufind.apps.coronawhy.org/vufind/Record/6g55l35h#Place651-7>	rdf:type	bf:Topic
7	<http://vufind.apps.coronawhy.org/vufind/Record/7bkf7wrs#Place651-7>	rdf:type	bf:Topic
8	<http://vufind.apps.coronawhy.org/vufind/Record/7hcr406k#Place651-7>	rdf:type	bf:Topic
9	<http://vufind.apps.coronawhy.org/vufind/Record/8pd6jdb5#Place651-7>	rdf:type	bf:Topic

Source: YASGUI

Do you know a lot of people that can use SPARQL?

PubChemRDF Use Cases: SPARQL query

PubChem

Q: What adverse effects of chemicals that are oral acute toxic according to GHS statement have been reported in PubMed literature, annotated by MeSH indexing?

```
PREFIX cito: <http://purl.org/spar/cito/>
PREFIX fabio: <http://purl.org/spar/fabio/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX meshv: <http://id.nlm.nih.gov/mesh/vocab#>
PREFIX mesh: <http://id.nlm.nih.gov/mesh/>

select distinct ?disease ?diseaselabel
where {
  ?compound sio:has-attribute/dcterms:subject/skos:broader/concept:Acute_Toxicity_Oral .
  ?syno sio:is-attribute-of ?compound .
  ?syno dcterms:subject ?meshconcept .
  ?pmid cito:discusses ?meshconcept .
  ?pmid fabio:hasSubjectTerm ?DQpair .
  ?DQpair meshv:hasQualifier mesh:Q000009 .
  ?pmid cito:discusses ?disease .
  ?disease rdf:type meshv:SCR_Disease .
  ?disease rdfs:label ?diseaselabel .
}
```

Source: Semantic annotation of the Laboratory Chemical Safety Summary in PubChem

CLARIAH conclusions

"*By developing these decentralised, yet controlled Knowledge Graph development practices we have contributed to increasing interoperability in the humanities and enabling new research opportunities to a wide range of scholars. However,* ***we observe that users without Semantic Web knowledge find these technologies hard to use, and place high value in end-user tools that enable engagement****. Therefore, for the future we emphasise the importance of tools to specifically target the goals of concrete communities – in our case, the analytical and quantitative answering of humanities research questions for humanities scholars. In this sense, usability is not just important in a tool context; in our view, we need to empower users in deciding under what models these tools operate.*" (CLARIAH: Enabling Interoperability Between Humanities Disciplines with Ontologies)

Chicken-egg problem: users are building tools without data models and ontologies but in reality they need to build a knowledge graph with common ontologies first!

Linked Data integration challenges

- datasets are very heterogeneous and multilingual
- data usually lacks sufficient data quality control
- data providers using different modeling schemas and styles
- linked data cleansing and versioning is very difficult to track and maintain properly, web resources aren’t persistent
- even modern data repositories providing only metadata records describing data without giving access to individual data items stored in files
- difficult to assign and manually keep up-to-date entity relationships in knowledge graph

CoronaWhy has too much information streams that seems to be impossible to integrate and give back to COVID-19 researchers. So, do we have a solution?

Bibliographic Framework (BIBFRAME) as a Web of Data

“The Library of Congress officially launched its Bibliographic Framework Initiative in May 2011. The Initiative aims to re-envision and, in the long run, implement a new bibliographic environment for libraries that makes "the network" central and makes **interconnectedness commonplace**.”

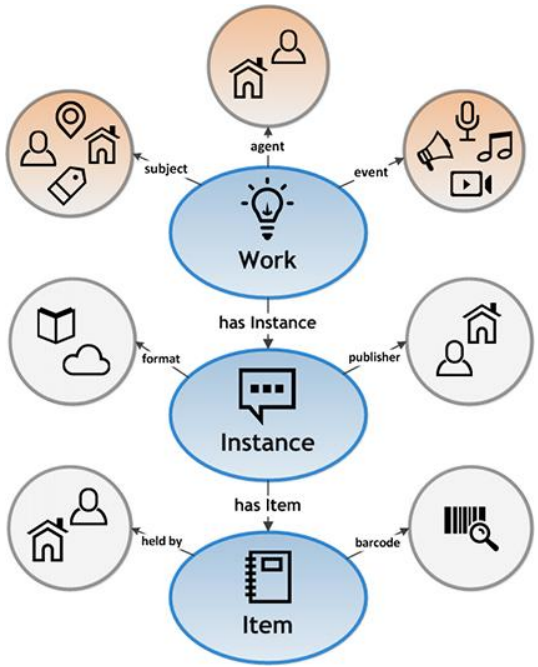
“Instead of thousands of catalogers repeatedly describing the same resources, the effort of one cataloger could be shared with many.” (Source)

In 2019 BIBFRAME 2.0, the Library of Congress Pilot, was announced.

Let’s take a journey and move from domain specific ontology to bibliographic!

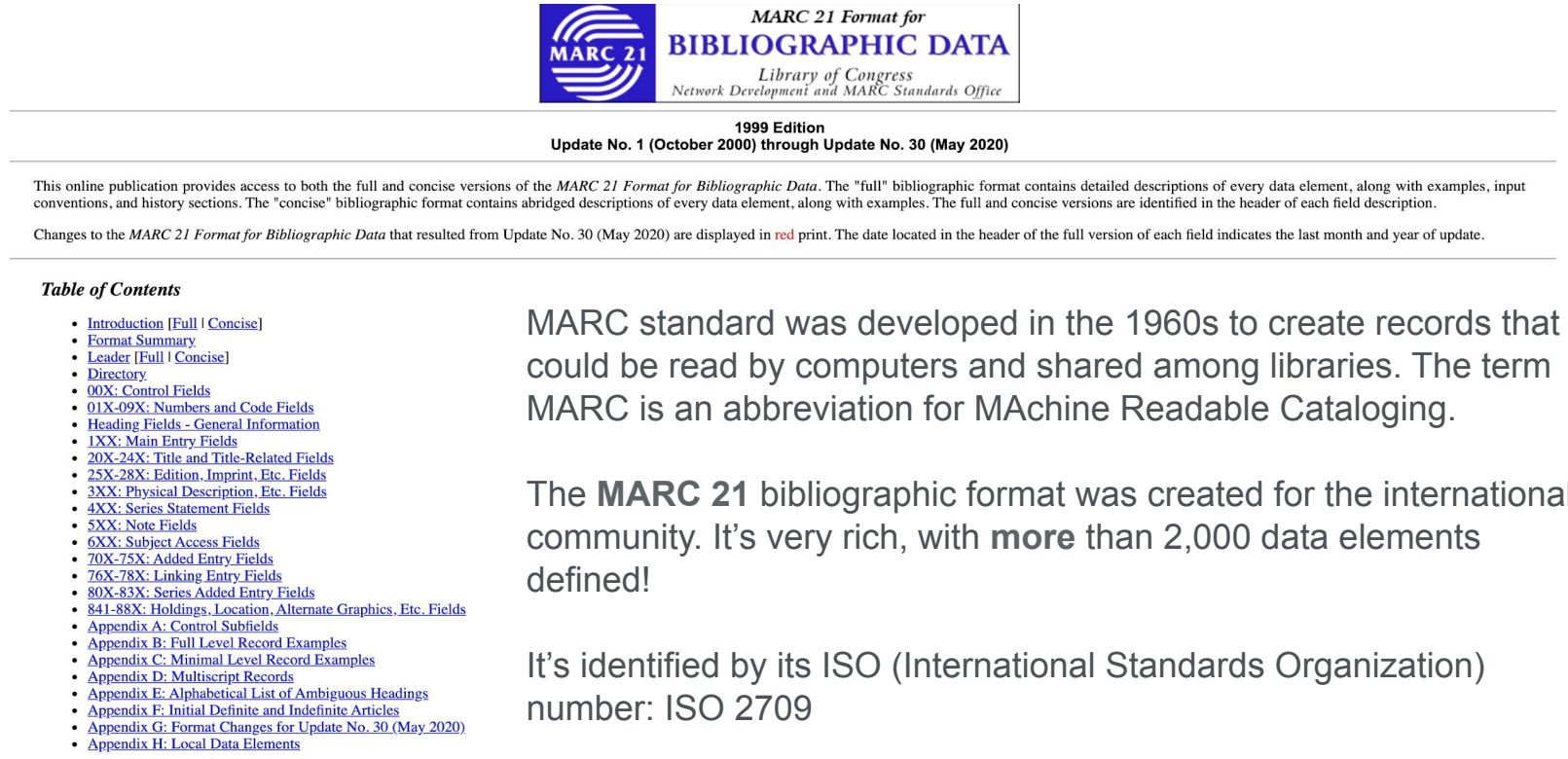
BIBFRAME 2.0 concepts

- **Work**. The highest level of abstraction, a Work, in the BIBFRAME context, reflects the conceptual essence of the cataloged resource: authors, languages, and what it is about (subjects).
- **Instance**. A Work may have one or more individual, material embodiments, for example, a particular published form. These are Instances of the Work. An Instance reflects information such as its publisher, place and date of publication, and format.
- **Item.** An item is an actual copy (physical or electronic) of an Instance. It reflects information such as its location (physical or virtual), shelf mark, and barcode.
- **Agents**: Agents are people, organizations, jurisdictions, etc., associated with a Work or Instance through roles such as author, editor, artist, photographer, composer, illustrator, etc.
- **Subjects**: A Work might be “about” one or more concepts. Such a concept is said to be a “subject” of the Work. Concepts that may be subjects include topics, places, temporal expressions, events, works, instances, items, agents, etc.
- **Events**: Occurrences, the recording of which may be the content of a Work.



Source: [the Library of Congress]

MARC as a foundation of the structured Data Hub



MARC 21 Format for
BIBLIOGRAPHIC DATA
Library of Congress
Network Development and MARC Standards Office

1999 Edition
Update No. 1 (October 2000) through Update No. 30 (May 2020)

This online publication provides access to both the full and concise versions of the *MARC 21 Format for Bibliographic Data*. The "full" bibliographic format contains detailed descriptions of every data element, along with examples, input conventions, and history sections. The "concise" bibliographic format contains abridged descriptions of every data element, along with examples. The full and concise versions are identified in the header of each field description.

Changes to the *MARC 21 Format for Bibliographic Data* that resulted from Update No. 30 (May 2020) are displayed in red print. The date located in the header of the full version of each field indicates the last month and year of update.

Table of Contents

MARC standard was developed in the 1960s to create records that could be read by computers and shared among libraries. The term MARC is an abbreviation for MAchine Readable Cataloging.

The **MARC 21** bibliographic format was created for the international community. It's very rich, with **more** than 2,000 data elements defined!

It's identified by its ISO (International Standards Organization) number: ISO 2709

Source: the Library of Congress, USA

How to integrate data in the common KG?

- Use MARC 21 as a basis for all bibliographic and authority records
- All controlled vocabularies should be expressed in MARC 21 format for Authority Data, we need to build an authority linking process with the “human in the loop” approach that will allow to verify AI predicted links.
- Different MARC 21 fields could be linked to the different ontologies and/or even interlinked. For example, we can get some entities linked to both MeSH and Wikidata in the same bibliographic record to increase the interoperability of the Knowledge Graph.
- Every CORD-19 paper can get a metadata enrichment provided by any research team working on the NLP extraction of entities, relations or linking CV together.

COVID-19 paper in MARC 21 representation

Holdings | Description | Comments | Similar Items | Staff View

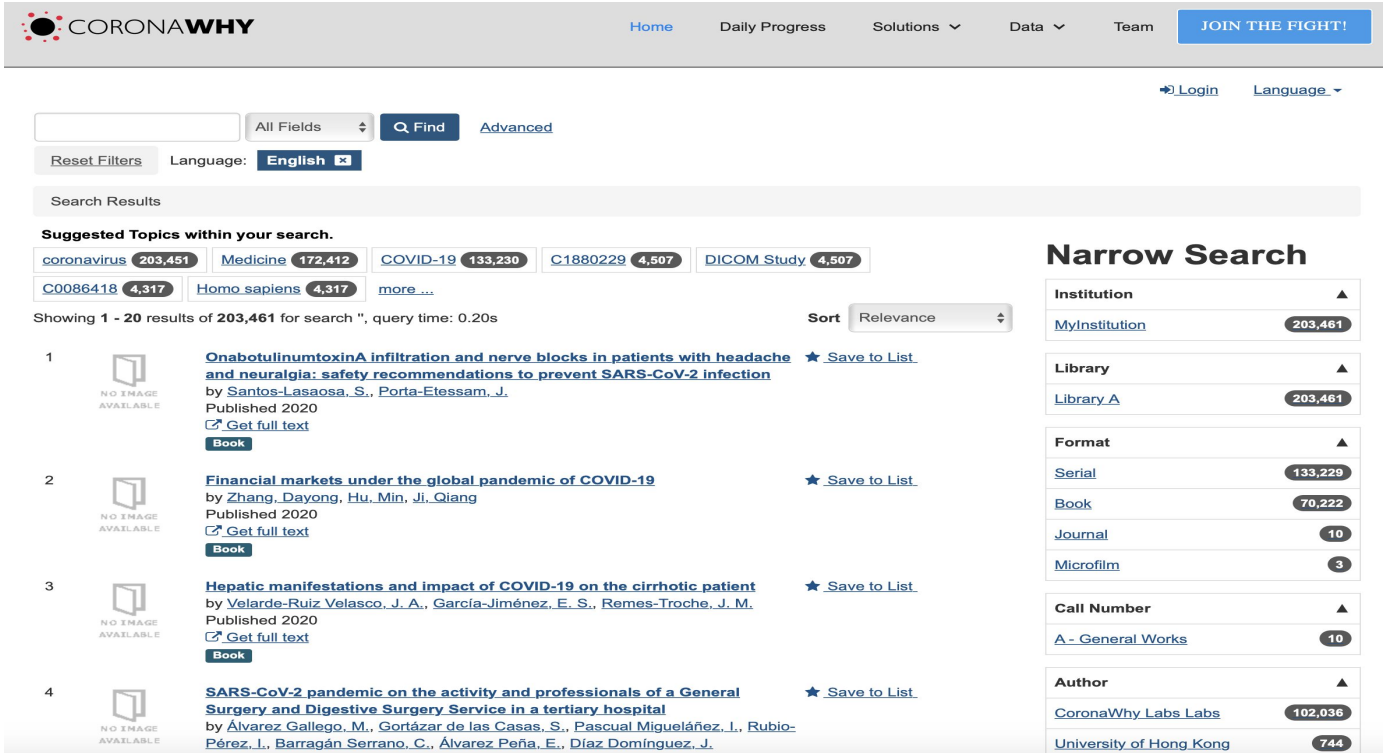
LEADER	06032cas a2201165 4500		
001	o32e9o0l		
008	//////cas//////cc ////////////////chi///		
245	1	0	|a The Transformation of Enterovirus Replication Structures: a Three-Dimensional Study of Single- and Double-Membrane Compartments
856	4	0	|u https://doi.org/10.1128/mbio.00166-11
651			|x coronavirus
651		4	|a Netherlands
370			|e Leiden, Netherlands
852			|a Department of Medical Microbiology, Molecular Virology Laboratory, Center of Infectious Diseases, Leiden University Medical Center, Leiden, The Netherlands
650	1	2	|a Act Relationship Type - transformation |x C1554215
650	1	2	|a Three-dimensional |x C0450363
650	1	2	|a Induce (action) |x C0205263
650	1	2	|a Membrane |x C0596901
650	1	2	|a Cells |x C0007634
650	1	2	|a Cell Microenvironment |x C1707328
650	1	2	|a RNA, Viral |x C0035736
650	1	2	|a Anabolism |x C0220781
650	1	2	|a structure |x C0678594
650	1	2	|a Enterovirus |x C0014383
650	1	2	|a Family Picornaviridae |x C0031886
650	1	2	|a Tissue membrane |x C0025255
650	1	2	|a Architecture as Topic |x C0003737

Structure of the bibliographic record:

- authority records contain information about authors and affiliation
- Medical entities extracted by NLP pipeline interlinked in 650x fields
- part of metadata fields generated and filled by Machine Learning models, part contributed by human experts
- provenance information kept in 833x fields indicating fully or partially machine-generated records
- Relations between entities stored in 730x fields

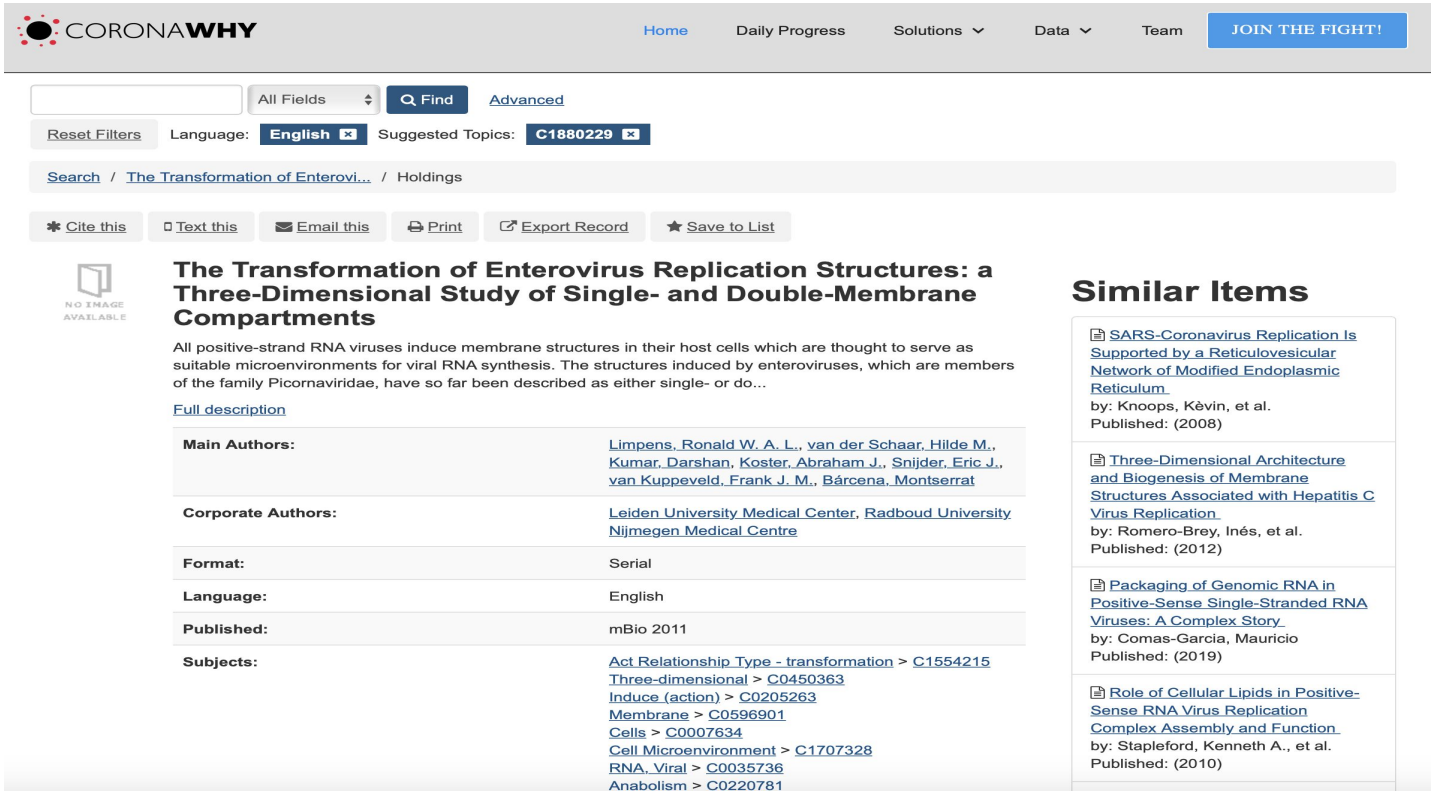
Source: CoronaWhy CORD-19 portal

Vufind discovery tool for libraries powered by MARC 21



Source: [CoronaWhy VuFind]

Landing page of publications from CORD-19

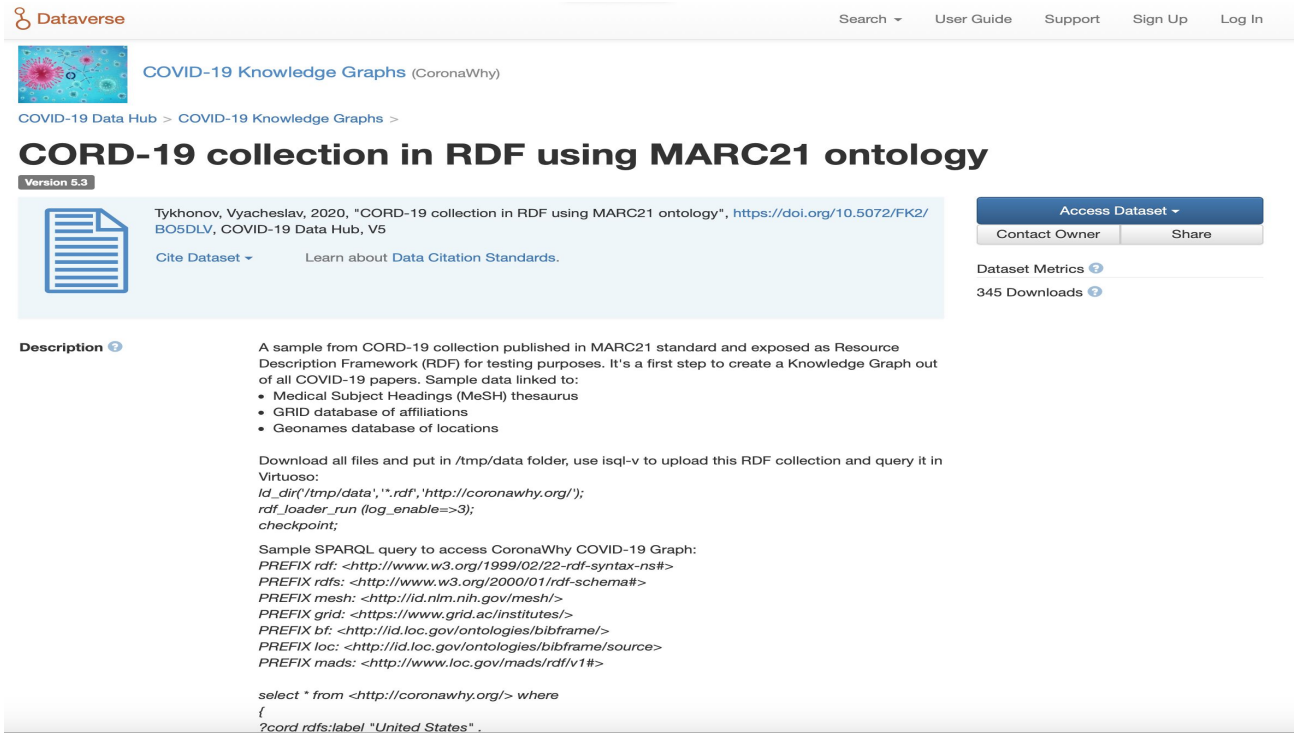


Source: [CoronaWhy VuFind]

CORD-19 collection in BIBFRAME 2.0

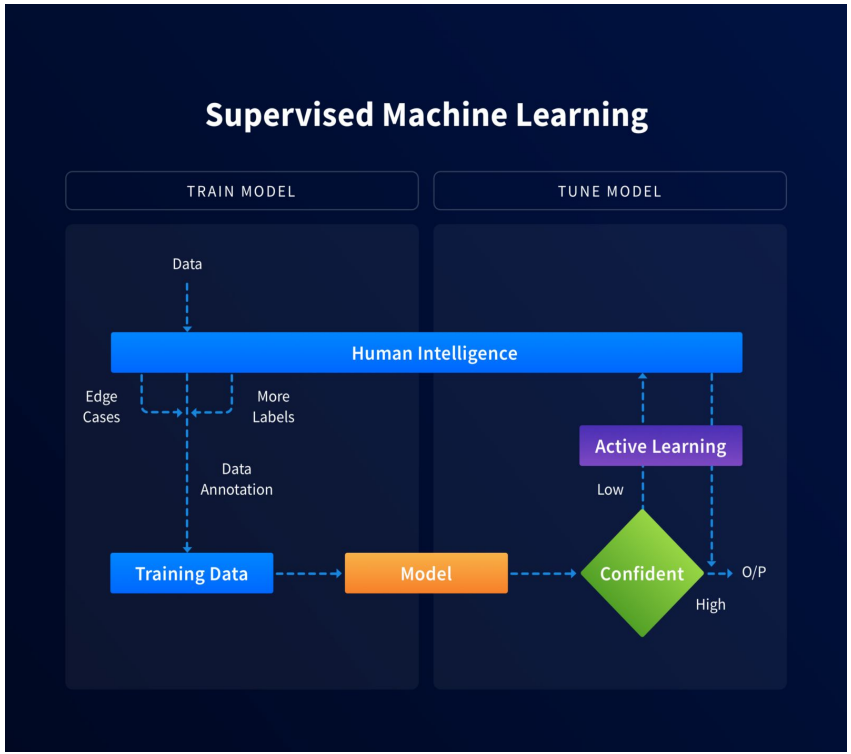
```
    <bf:Content rdf:about="http://id.loc.gov/vocabulary/contentTypes/txt">
      <rdfs:label>text</rdfs:label>
    </bf:Content>
  </bf:content>
  <bf:language>
    <bf:Language rdf:about="http://id.loc.gov/vocabulary/languages/%2Fch"/>
  </bf:language>
  <rdfs:label>The SARS Coronavirus S Glycoprotein Receptor Binding Domain: Fine Mapping and Functional Characterization</rdfs:label>
  <bf:title>
    <bf:Title>
      <rdfs:label>The SARS Coronavirus S Glycoprotein Receptor Binding Domain: Fine Mapping and Functional Characterization</rdfs:label>
      <bflc:titleSortKey>The SARS Coronavirus S Glycoprotein Receptor Binding Domain: Fine Mapping and Functional Characterization</bflc:titleSortKey>
      <bf:mainTitle>The SARS Coronavirus S Glycoprotein Receptor Binding Domain: Fine Mapping and Functional Characterization</bf:mainTitle>
    </bf:Title>
  </bf:title>
  <bf:subject>
    <bf:Place rdf:about="http://vufind.apps.coronawhy.org/vufind/Record/ofx0hvvs#Place651-6">
      <rdf:type rdf:resource="http://www.loc.gov/mads/rdf/v1#ComplexSubject"/>
      <rdfs:label>coronavirus</rdfs:label>
      <madsrdf:authoritativeLabel>coronavirus</madsrdf:authoritativeLabel>
      <madsrdf:componentList rdf:parseType="Collection">
        <madsrdf:Topic>
          <madsrdf:authoritativeLabel>coronavirus</madsrdf:authoritativeLabel>
        </madsrdf:Topic>
      </madsrdf:componentList>
    </bf:Place>
  </bf:subject>
  <bf:subject>
    <bf:Place rdf:about="http://vufind.apps.coronawhy.org/vufind/Record/ofx0hvvs#Place651-7">
      <rdf:type rdf:resource="http://www.loc.gov/mads/rdf/v1#Geographic"/>
      <rdfs:label>United States</rdfs:label>
      <madsrdf:authoritativeLabel>United States</madsrdf:authoritativeLabel>
    </bf:Place>
  </bf:subject>
  <bf:subject>
    <bf:Topic rdf:about="http://id.nlm.nih.gov/mesh/C1175175">
      <rdf:type rdf:resource="http://www.loc.gov/mads/rdf/v1#ComplexSubject"/>
      <rdfs:label>Severe Acute Respiratory Syndrome--C1175175</rdfs:label>
      <madsrdf:authoritativeLabel>Severe Acute Respiratory Syndrome--C1175175</madsrdf:authoritativeLabel>
      <madsrdf:componentList rdf:parseType="Collection">
        <madsrdf:Topic>
          <madsrdf:authoritativeLabel>Severe Acute Respiratory Syndrome</madsrdf:authoritativeLabel>
        </madsrdf:Topic>
        <madsrdf:Topic>
          <madsrdf:authoritativeLabel>C1175175</madsrdf:authoritativeLabel>
        </madsrdf:Topic>
      </madsrdf:componentList>
      <bf:source>
        <bf:Source>
          <bf:code>mesh</bf:code>
        </bf:Source>
      </bf:source>
    </bf:Topic>
  </bf:subject>
  <bf:subject>
    <bf:Topic rdf:about="http://id.nlm.nih.gov/mesh/C1283195">
      <rdf:type rdf:resource="http://www.loc.gov/mads/rdf/v1#ComplexSubject"/>
      <rdfs:label>Mapping (action)--C1283195</rdfs:label>
      <madsrdf:authoritativeLabel>Mapping (action)--C1283195</madsrdf:authoritativeLabel>
      <madsrdf:componentList rdf:parseType="Collection">
        <madsrdf:Topic>
          <madsrdf:authoritativeLabel>Mapping (action)</madsrdf:authoritativeLabel>
        </madsrdf:Topic>
```

CoronaWhy Graph published as RDF



Source: [CoronaWhy Dataverse]

Human-in-the-Loop for Machine Learning



Source: Hackernoon.com

"Computers are incredibly fast, accurate and stupid; humans are incredibly slow, inaccurate and brilliant; together they are powerful beyond imagination."
Albert Einstein

"A combination of AI and Human Intelligence gives rise to an extremely high level of accuracy and intelligence (Super Intelligence)"

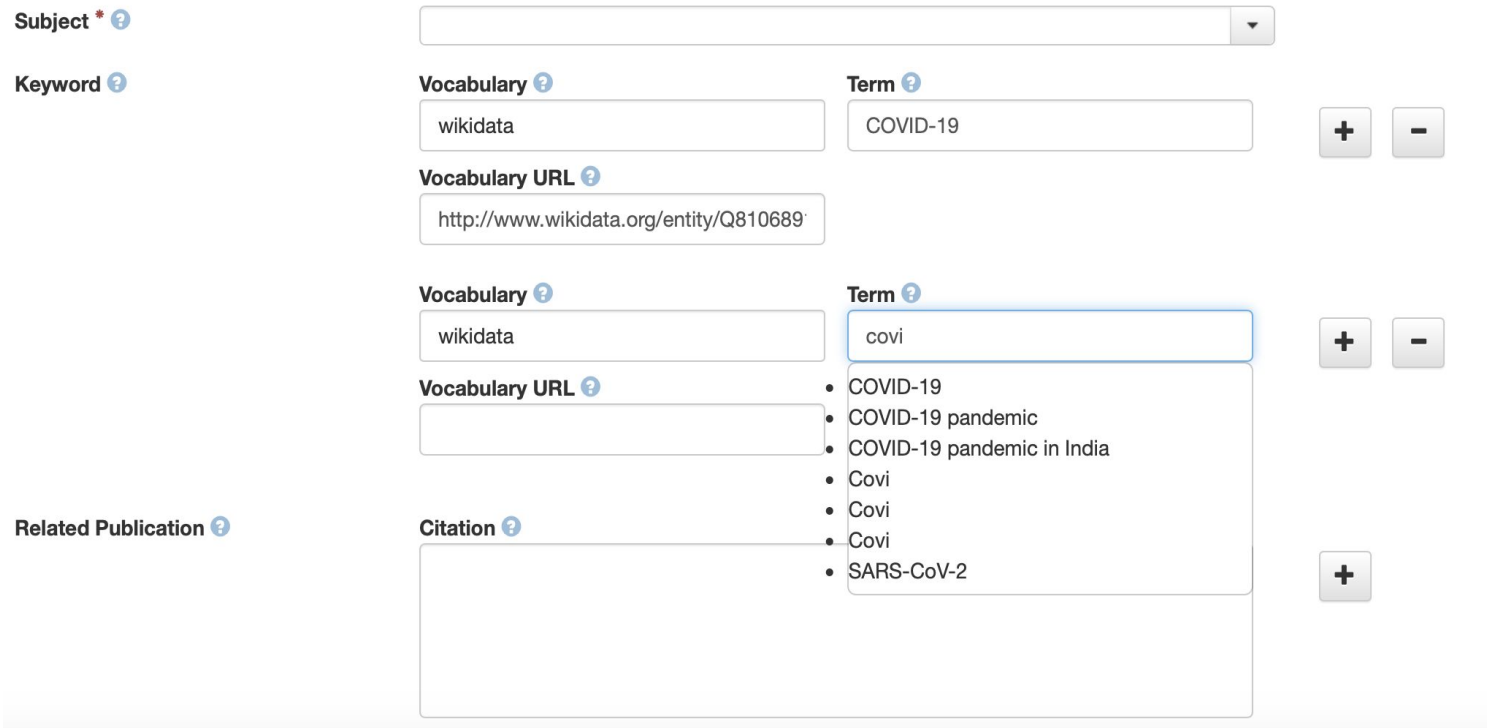
CLARIAH and Network Digital Heritage (NDE) GraphiQL

CoronaWhy is running own instance of [NDE on Kubernetes] cluster and maintains the support of another ontologies (MeSH, ICD, NDF) available via their SPARQL endpoints!

If you want to query API:

curl "http://nde.dataverse-dev.coronawhy.org/nde/graphql?query=%20%7B%20terms(match%3A%22COVID%22%2Cdataset%3A%5B%22wikidata%22%5D)%20%7B%20dataset%20terms%20%7Buri%2C%20altLabel%7D%20%7D%20%7D"

Increasing Dataverse metadata interoperability

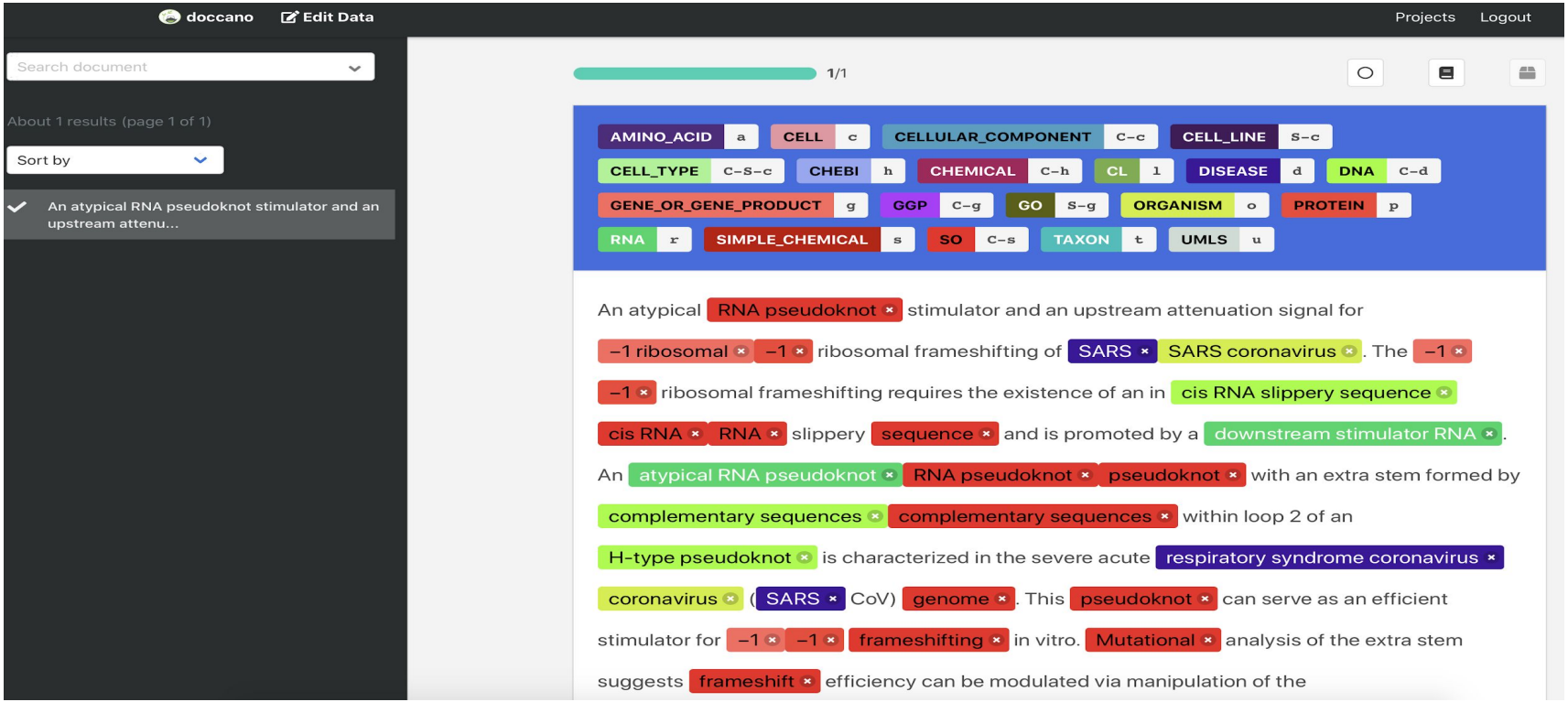


External controlled vocabularies support contributed by SSHOC project (data infrastructure for the EOSC)

Hypothes.is annotations as a peer review service

1. AI pipeline does domain specific entities extraction and ranking of relevant CORD-19 papers.
2. Automatic entities and statements will be added, important fragments should be highlighted.
3. Human annotators should verify results and validate all statements.

Doccano annotation with Machine Learning



Source: Doccano Labs

Building an Operating System for Open Science

CoronaWhy Common Research and Data Infrastructure is distributed and robust enough to be scaled up and reused for other tasks like cancer research

All services are build from Open Source components

Data processed and published in FAIR way, the provenance information is the part of our Data Lake

Data evaluation and credibility is the top priority, we're providing tools for the expert community for the verification of our datasets

The transparency of data and services guarantees the reproducibility of all experiments and get bring new insights in COVID-19 research

CoronaWhy Common Research and Data Infrastructure

Data preprocessing pipeline implemented on Jupyter notebook Docker with extra modules.

Dataverse as data repository to store data from automatic and curated workflows.

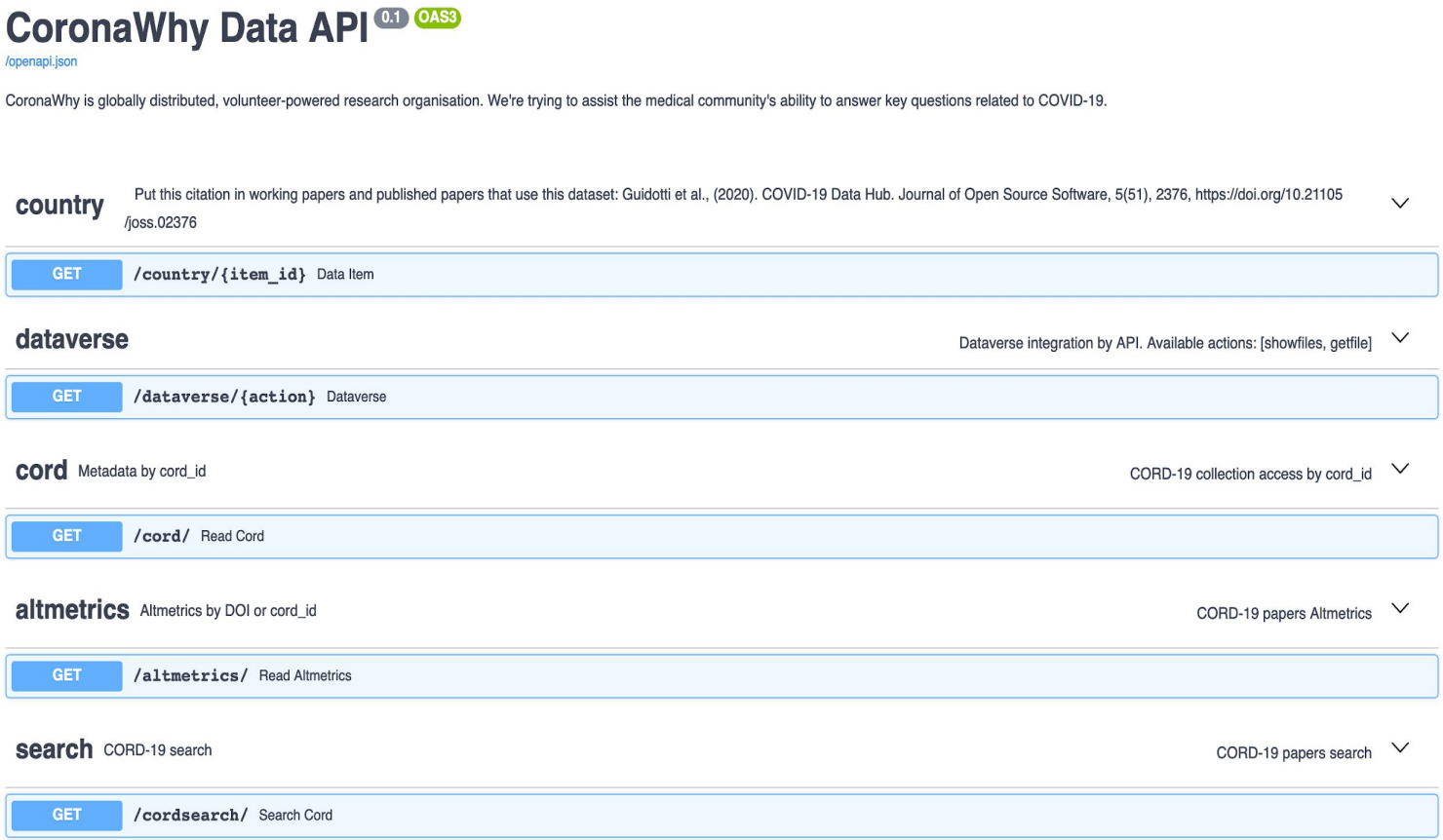
Elasticsearch has CORD-19 indexes on sections and sentences level with spacy enrichments. Other indexes: MeSH, Geonames, GRID.

Hypothesis and Doccano annotation services to annotate publications

Virtuoso and GraphDB with public SPARQL endpoints to query COVID-19 Knowledge Graph

Other services: Colab, MongoDB, Kibana, BEL Commons 3.0, INDRA, Geoparser, Tabula

https://github.com/CoronaWhy/covid-19-infrastructure



Source: [CoronaWhy API](#) built on FastAPI framework for Python

Thank you! Questions?

@Slava Tykhonov on CoronaWhy Slack

vyacheslav.tykhonov@dans.knaw.nl

www.coronawhy.org

www.dans.knaw.nl/en